DOCUMENT RESUME

ED 423 282 TM 029 087

AUTHOR Holweger, Nancy; Taylor, Grace
TITLE Differential Item Functioning by Gender on a Large-Scale Science Performance Assessment: A Comparison across Grade Levels.
PUB DATE 1998-00-00
NOTE 31p.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; *Elementary School Students; Grade 5; Grade 8; *Item Bias; *Junior High School Students; *Performance Based Assessment; Performance Factors; *Science Tests; *Sex Differences; State Programs; Testing Programs

ABSTRACT

The fifth-grade and eighth-grade science items on a state performance assessment were compared for differential item functioning (DIF) due to gender. The grade 5 sample consisted of 8,539 females and 8,029 males and the grade 8 sample consisted of 7,477 females and 7,891 males. A total of 30 fifth grade items and 26 eighth grade items were examined for DIF using the Multilog software package. There was substantial DIF on certain items: in the grade 5 data, four items exhibited the largest DIF. Females performed better on these items, none of which involved using an algorithm. Interestingly three of the four items with large DIF were in the section on salinity, a topic that perhaps is differentially interesting to males and females. In addition, these four items were embedded in real contexts, and each involved considerable writing. Results also show that there is considerably less DIF in the grade 5 science items than in the grade 8 science items. DIF indices for grade 8 range from 0.02 to 0.56, while for grade 5 they range from 0.02 to 0.33. Adolescence, which affects eighth graders more strongly than fifth graders, may emphasize gender differences because of the development of secondary sex characteristics. Implications of these findings are discussed. (Contains 32 references, two tables, and four figures.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 423 282

Differential Item Functioning by Gender on a Large-Scale Science Performance Assessment : A Comparison Across Grade Levels

by

Nancy Holweger
and
Grace Taylor

TM029087

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Nancy Holweger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Introduction

The use of performance assessments in state testing has grown increasingly popular over the last five years (State Student Assessment Programs Database, Council of Chief State School Officers and North Central Regional Education Laboratory, 1996). In a 1996 survey of state assessment directors, 39 states reported using non-multiple choice assessment exercises (p. 105). Out of the 11 remaining states who currently do not use non-multiple choice assessments, 6 states plan to develop non-multiple choice assessments in the future (CCSSO & NCREL, p. 105).

At the same time that state testing directors have begun to explore the uncharted world of performance assessment, they have also faced increasing pressure from state legislatures and the public to use testing for accountability purposes. Florida, for example, uses a writing assessment to help identify low performing schools (CCSSO & NCREL, p. 271). Kentucky also uses performance assessments for school accountability purposes (CCSSO & NCREL, p. 262). The push for performance assessment, coupled with the push for high-stakes uses has caused concern among the measurement community (Linn, 1993; Linn, Baker & Dunbar, 1991; Shavelson, Baxter & Pine, 1992). In particular, measurement experts have raised questions about the validity and generalizability of some states' performance measures (Bob Linn, personal communication, April 12, 1997). The KIRIS Technical Report discusses this issue in some detail. Coupled with concern about the overall validity of performance measures, is a concern about bias. Do performance measures favor one group over another? This is an example of the type of question being asked. If performance measures are going to be used for accountability purposes, it is crucial that they fairly represent the achievement of all students who take the test.

Even if performance measures seem to favor one group over another, it does not necessarily mean that a test or test item is biased. The gap in performance could also be due to real differences between the groups. In order to get at bias, extensive studies of the test and individual items must be completed. One indicator of possible item bias is differentjal item functioning (DIF). DIF procedures allow the researcher to identify items that function differently for various groups of examinees.

The purpose of this paper is to compare the fifth grade and eighth grade science items on a State Performance Assessment for differential item functioning due to

gender. Our research questions are as follows:

1. Is there differential item functioning due to gender on the fifth grade science items of the 1994 State Performance Assessment? Is there differential item functioning due to gender on the eighth grade science items of the 1994 State Performance Assessment? In what direction and to what degree does DIF operate on these items?
2. How does the DIF of the fifth grade science items compare to the DIF of the eighth grade science items?

Literature Review

In order to understand the context of the present study, it is necessary to review some background information. There are two major areas of research that we need to consider in this review. One of these areas is the theory and method of differential item functioning. The other area is gender differences in assessment, particularly, in science assessment.

DIFFERENTIAL ITEM FUNCTIONING

The first topic to be considered in this review of the literature on differential item functioning is the criterion by which DIF is operationalized. Certain DIF analysis techniques use an external criterion and others use an internal criterion. Those techniques that rely upon external criteria are based upon the predictive validity paradigm. Cleary (1968; cited in Camilli & Shepard, 1994) proposed that differential prediction of the criterion could be a source of evidence in the determination of differential item functioning. According to Cleary, groups of test takers should have the same regression equations to predict the relationship between the independent and dependent variables. Different prediction equations are evidence of differential prediction from the total test and possible bias. Thorndike's constant ratio model and the "performance fair" models by GATB are alternative or competing models to Cleary's original formulation (Camilli & Shepard, 1994). The advantage of using regression equations to measure DIF is that they are external to the test. There are concerns with the technical quality of these techniques, however (Camilli & Shepard, 1994).

The internal criterion measures of DIF are based upon item level differences. These measures examine the extent to which individual items operate differently for

subgroups of examinees. Item differences are considered relative to other items on the same test. It is critical, when using these measures of DIF, to logically examine the reported differences in order to make judgments about the real differences in subgroups vs. the differences due to the "bias" of the items. If the DIF is a function of real group differences on the construct, it is not an indicator of item "bias" (Camilli and Shepard, 1994).

It is important to reiterate that, although DIF is a powerful indicator of multidimensionality in a test, it may miss item bias in cases where the test is systematically biased or when the majority of the test items are biased. Differential item functioning procedures are also of importance because they force test developers to consider whether the items are functioning as intended. Camilli & Shepard (1994) argue that caution should be exerted, however, when using differential item functioning to determine the use of a test in a specific situation. Test validity or fairness can only be assessed within a larger framework.

The second topic to be considered in this review of the literature on differential item functioning is that of the method of analysis. In item response theory, items are considered invariant across samples of examinees. This allows items to be examined for DIF. If the item characteristic curves (ICCs) are different, after common scaling, for two or more groups, the item is said to have DIF. Several techniques are used to consider the differences in ICCs and item parameters. These techniques fall into the broad categories of indices and statistical tests of DIF (Camilli & Shepard, 1994).

The index techniques include "signed" and "unsigned" methods. If the DIF is uniform at all ability levels, the ICCs will be more or less parallel and the area between the curves can serve as a measure of DIF. If the DIF is non-uniform and the ICCs cross, the item will appear more difficult for one group at low ability and for the other group at high ability. If the "unsigned" method is used with items of non-uniform DIF, the index will underestimate the actual DIF. Examination of the corresponding ICCs is necessary to reveal this potential problem (Hambleton & Swaminathan, 1985). Such an examination of the ICCs for this project indicates that they are primarily uniform.

The area between the ICCs can be integrated across all levels of theta in a simple index or it can be weighted for number of examinees at various points along the theta scale. Our study employs the simple, unweighted calculation of DIF indices, because it is more straightforward to apply and therefore more practical for this project.

Other IRT techniques such as comparison of item parameters and comparison of fit

have been used as measures of DIF. In the case of high-stakes DIF analysis, it has been suggested that both an area difference method and a comparison of item parameters be used to search for item bias. The method employed in our study does indeed include the calculation of an area difference and the comparison of the item parameters (a and b) to ascertain the DIF.

GENDER DIFFERENCES IN ASSESSMENT

Gender Differences in Assessment - General

Several studies have considered the differences in assessment items between the genders. These studies have identified aspects of test items that differentially favor males or females. They have also addressed the issue of item format in regard to gender differences. We will briefly review the findings of these studies to provide additional background for our current research project.

The majority of the recent research on gender differences in assessment has focused upon mathematics assessments. These various studies (Berberoglu, 1995; Ryan & Chiu, 1996; Lane et al., 1995; Wang & Lane, 1994; Friedman, 1996; Catsambis, 1994) all reach a similar conclusion about the differential item functioning due to gender on mathematics test items. These authors all agree that females perform better on mathematical algorithm problems and males perform better on mathematical reasoning problems.

Another significant difference in the performance of males versus females on assessment items is based upon the interest level of the items. Westers and Kelderman (1992) found that the level of interest (attractiveness) of various distracter items was different for the two genders and their performance levels depended upon how many items were attractive to each sex. This evidence supports the intuitively logical concept that items of high interest to one or the other of the sexes will show differential item functioning due to gender. Item bias of this type seems to be relatively easily detected, however. Consequently, it may not be a major factor in well-designed assessments.

Context is also a factor in the gender differences on assessments. Research indicates that females perform better than males on items with embedded context. This finding has been replicated by Bransky and Qualter (1993) with physics assessments, Gallagher (1992) with the SAT math test, and Jackson et al. (1995) with numerical tests. Context is related to the issue of item format that has received some

attention in the research literature of late.

There is some limited evidence that item format differentially favors males and females. Males perform relatively better on multiple choice tests and females tend to do better on essay tests (Bolger & Kellaghan, 1990; Breland, 1991; Bridgeman & Lewis, 1994). Consequently, it is hypothesized that females perform relatively better than males in assessments that require more writing. This hypothesis is corroborated by evidence from several other studies (Severiens, Tem & Geert, 1994; Lawrence & Curley, 1989; Carlton & Harris, 1992).

The format of most performance assessments focuses upon written responses. (Note that if the assessment used only written responses this might disappear as DIF.) There is a real possibility that females will be favored in performance assessments because of two factors. These factors have been alluded to above. Performance assessments are more context specific. Also, performance assessments are more writing dependent. Jovanovic et al. discuss this possibility in their 1994 article. Also, Marion and Shepard, 1995, found that performance assessments favor females in their study of a large scale performance assessment from Maryland. This is the same dataset used in the present study. They found that girls scored higher than boys in all six subjects on the test battery in each year and for all three grade levels (grade 3, 5 and 8).

Another factor that may possibly affect differential item functioning due to gender is that males seem to outperform females on items that are visually/spatially oriented. There is no research confirmation of this hypothesis at this time. However, the authors of this paper noticed this trend and, the current study may further illuminate this possible source of item bias.

Gender Differences in Assessments - Science

In addition to the gender differences in assessments that were discussed in the previous section, there are some indications that, within the domains of science content, gender is also a factor. Linn and Hyde (1989) found small but consistent gender differences favoring males in knowledge of science content. However, when patterns of difference were examined by content area, girls performed as well or better than boys in the life sciences and scientific inquiry, but males consistently outperformed females in the physical sciences and, to a lesser degree, the earth sciences.

The three basic types of science assessment items are information items, process items and the fairly new category of science, technology and society items. There are no gender differences when science process and science knowledge items are examined in isolation from one another. There is an exception to this rule. Proportional reasoning questions, a component of mathematical reasoning, favor males. More importantly, when science process questions are combined with science content questions, the sex differences favoring males are magnified (Linn & Hyde, 1989). This may contribute to the previously reported higher scores for males on physical science items. Physical science lends itself more to problem solving and application questions.

Method

Data Source and Sample

To perform our DIF analysis, we used the 1994 science test data from a State Performance Assessment - grade 5 and grade 8. After removing students with total zero scores, our grade 5 sample consisted of 8, 539 females and 8, 029 males and our grade 8 sample consisted of 7,477 females and 7,891 males. We removed students who scored a zero on all items because the program we used, Multilog (SSI, 1991), does not function with total zero scores.

Structure of the Grade 5 Assessment

The State Performance Assessment for fifth grade science consists of a variety of integrated performance tasks designed to elicit higher-order thinking skills from the fifth graders. The tasks are grouped in four to six test booklets and a matrix sample is used to collect a variety of school level information. Items draw from the following domains:

1. Concepts of Science: This area deals with unifying themes from life, physical, earth and space science.
2. Nature of Science: This area asks students to explain and interpret information about scientific phenomena.
3. Habits of the Mind: This area asks student to demonstrate ways of thinking about science.

4. Attitudes: Not assessed.

5. Science Processes: This area asks students to use language, instruments, methods and materials to collect, organize and interpret information.

6. Application of Science: This area asks students to apply what they have learned in science to solve problems.

While the fifth grade science test is primarily a test of science, many of the items on the test are a mixture of science and reading or science and math. Listed below are the components included in each test item.

Items with Science Only

Items #6, 7, 8, 12, 14, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30

Items with Science and Reading

Items # 1, 2, 3, 4, 5

Items with Science and Math

Items # 9, 10, 11, 13, 15, 16, 17

The 8th grade science test items are also mixtures of these components. However, specifications concerning these components are not available for the 8th grade science data from this 1994 assessment.

Scoring and Calibration

The scoring of student responses was conducted in a manner similar to the scoring of many other forms of open-ended assessments. The scoring process for the assessment included the development of a scoring guide (rubric), extensive training of scorers that required scorers to meet a minimum standard on a qualifying set of papers (70-80% exact agreement, depending on the type of task), and a series of quality control procedures including read-behinds, check-sets, and reader-effects reliability studies. To calibrate the item raw scores a graded response model was used for converting the responses to these items into scale scores (Yen, et al., 1992).

Analysis

Thirty items from the 1994 Science Performance Assessment for 16,568 fifth grade students were examined for differential item functioning (DIF) between males and females. Twenty-six items from the same assessment for 15,369 eighth grade students were also examined. Females were chosen as the reference group because

the female group displayed higher raw score performance on the majority of items as well as on the overall test.

Because some of the items on the assessment have more than two categories, (for example, in the grade 5 assessment, 18 items had three categories and 1 item had four categories), the items were examined using Mulitlog (SSI, 1991), a software package designed to conduct IRT analyses of items with multiple response categories. The item responses were first analyzed using the male and female responses in the same data set. Item parameters were found using Samejima's (1968, cited in Thissen, 1991) graded response model with a random maximum marginal likelihood estimation procedure. Multiple b parameters were found for items with more than one category. The item parameters from the random MML procedure were then fixed and individual theta scores were estimated for each student using the same data set. The separate means and standard deviations for the theta scores of males and females were calculated from the individual theta scores.

Male and female item parameters and individual theta estimates were then derived for each group using separate data sets. An identical procedure (graded response, MML and fixed estimation) was used to find item parameters and theta estimates for the separate male and female groups. The item parameters from the separate groups were placed on the same scale using the mean and standard deviations from the separate male and female runs, and the means and standard deviations from the males and females in the full group run.

The following linear transformation was used to put the item and ability parameters on comparable scales.

$$\theta = c + d\theta^*$$

$$b = c + b^* (d)$$

$$a = a^*/d$$

The c and d constants were calculated as follows:

$$c = \theta - d\theta^*$$

$$d = S\theta/S\theta^*$$

Where θ, $S\theta$, b, and a equal the group mean, standard deviation, and b and a parameters of (either) the males or the females in the joint data set, and θ^*, $S\theta^*$, b^*, and a^* equal the group mean, standard deviation, b and a parameters of (either) the males or the females in one of the separate data sets.

For grade 5, item characteristic curves for each item and each group with a DIF of

.15 or higher were plotted using the scaled item parameters. Those items with a DIF of .30 or higher on the grade 5 assessment have been included in Appendix A. The following probability index suggested by Linn et al. (1981, cited in Hambleton & Swaminathan, 1985) was used to find the unsigned area between the item characteristic curves of the males and females.

$$DIF = \sum \{[P_{i1}(\theta k) - P_{i2}(\theta k)]^2 \Delta\theta\}^{1/2}$$

Results

Items

There are several problems with the design of this assessment that might profoundly affect the DIF analysis. The first problem with this assessment is that it most likely measures more than one construct. As we pointed out in the methods section of this report, five of the thirty items in the grade 5 assessment were designed to assess science and reading concurrently and seven of the items were designed to assess science and math concurrently. The remaining eighteen items were "pure" science items. One of the necessary assumptions in item response theory and its extension in differential item functioning analysis is that the construct being measured is unidimensional (Hambleton & Swaninathan, 1985). Clearly, the assessment is not unidimensional.

There is additional evidence for the multidimensionality of this measure. We pointed out in the methods section of this paper that there are six domains of science upon which the assessment draws. These domains include concepts of science, nature of science, habits of the mind, attitudes, science processes and applications of science. Different items in the assessment emphasize varying combinations of these constructs. We see, once again, that this assessment is not unidimensional.

The difficulties with multidimensionality in item response theory have been recently addressed by Douglas, Roussos and Stout, (1996). In their 1996 article in the Journal of Educational Measurement, they discuss the multidimensionality assumption that is such a problem in this measurement instrument. They propose suggestions to resolve this problem. It is beyond the scope of this project to attempt multidimensional IRT analysis. However, future research on our data set should include the consideration of these suggestions.

The second problem with the aplication of IRT to evaluate DIF in this assessment is

that it also violates the item response theory assumption of local independence (Hambleton & Swaminathan, 1985). For example, in the grade 5 assessment, items #1 through #8 and #18 all pertain to soil tests. Items #9 through #17 are concerned with levers. Items #19 through #30 all relate to salinity. When items are related to one another like this, they cannot actually be considered separate items in the item response theory sense. Again, Douglas, Roussos and Stout address this issue in their 1996 article. They suggest the analysis of item "bundles" or "testlets" as a technique for overcoming this difficulty. In the Maryland assessment, for instance, we can use their suggestions to analyze the "soil test", "levers" and "salinity" testlets. It is beyond the scope of this project to do such an analysis. However, it would be an appropriate direction for additional research.

The third and final problem with the design of the assessment is that the students work independently at times and, at other times, work in pairs or groups of four. The measures of individual ability that are produced by an item response theory analysis of this data are highly suspect. These measures reflect some individual ability and some collaborative ability or even "acquired" ability. It has been our experience that students working in groups often produce work that is most reflective of the ability of those with the most profound understanding of the material and processes. It might be possible to treat group or pair collaborative items as test "bundles", as suggested by Douglas, Roussos and Stout (1996). This collective analysis can be conducted based upon the same assumptions as those for items that are connected on the basis of content. However, the measures of individual ability that might be produced with this technique are still completely questionable.

Item response theory analysis of the State Performance Assessment is limited by three different difficulties. The IRT assumptions of unidimensionality and local independence are violated in this assessment. Also, the group collaboration on many of the items confounds the measurement of individual ability. Techniques that might overcome the first two limitations have been proposed but are beyond the scope of this project.

DIF FINDINGS

Table 1

The findings of our differential item functioning analyses of each of the thirty questions in the fifth grade science, 1994, Performance Assessment can be found in

Table 1. The findings of our differential item functioning analyses of each of the twenty-six questions in the eighth grade science, 1994, Performance Assessment can be found in Table 2. For grade 5, the range of DIF found for the b1 indices (the DIF between rating categories 0 and 1) is from .03 to .33. The range of DIF found for the b2 indices (the DIF between rating categories 1 and 2) is from .02 to .19. The DIF for the b3 index (the DIF between rating categories 2 and 3) is .05. For grade 8, the range of DIF for the b1 indices is .04 to .56 and the range of the DIF found for the b2 indices is .02 to .54. The DIF for the b3 index is .13.

For grade 5, those items that were identified as being a combination of science and reading constructs have DIF indices that range from .16 to .27 in the b1 category and from .11 to .19 in the b2 category. In comparison to the range of the entire set of items, these DIF ranges are moderate. Those items that were identified as being a combination of science and math constructs have DIF indices that range from .03 to .17 in the b1 category and from .03 to .12 in the b2 category. These DIF ranges are very low compared to the range of the entire set of items. The strictly science items range from .08 to .33 in the b1 category, from .02 to .17 in the b2 category and are .05 in the b3 category. These items have the highest DIF ranges in the item set. A similar composition of items are not available for the grade 8 sample.

The implications of this trend are interesting. It appears that gender differences in performance assessment items are the greatest in science only items, less in science with reading items and the least in science with math items. It is possible that, in performance assessments, adding a second dimension that favors one or the other gender, actually suppresses the differences due to the different science domains.

Those items that belong to the first "bundle" or "testlet" of items related to soil testing have a range of DIF from .13 to .32 in the b1 category and from .03 to .19 in the b2 category. The items that belong to the levers testlet range from .03 to .17 in the b1 category, from .03 to .12 in the b2 category and are .05 in the b3 category. The items that belong to the salinity testlet range from .09 to .33 in the b1 category and from .02 to .17 in the b2 category. These ranges indicate that the levers testlet shows the least DIF, the soil test testlet shows a moderate amount of DIF and the salinity testlet shows the most DIF. These testlets roughly correspond to the items with added reading and math dimensions that are discussed above, however. Consequently, this trend may be an artifact of the multidimensionality issue rather than the local independence issue.

Appendix A

The ICCs for each of the items for grade 5 with a DIF index above .30 are located in Appendix A. They are provided for the purpose of examining the "signed" versus "unsigned" aspect of differential item functioning. If the ICCs cross, the signed but not the unsigned DIF is suppressed because the better performance for one group at the low end of the theta scale is canceled out by the better performance of the other group at the high end of the theta scale (Hambleton & Swaminathan, 1985). In our study, we found a small degree of this cross-over in items #17, 27 and 28 and a moderate degree of it in item #24. The relatively moderate DIF indices for these items are each artificially low due to the fact that we used an unsigned area measure to calculate the DIF. However, the ICCs for each of the items in Appendix A, those items with substantial DIF, are not largely affected by this suppression. Thus, those items with large DIF are adequately calculated with the "unsigned" formula for differential item functioning. The previously discussed trends in the results are not affected by this suppression, either. ICCs for grade 8 are not provided.

The grade 5 ICCs can also be examined for difficulty. It is clear that items #18, 20, 22 and 23 are considerably more difficult than the rest of the items in this assessment. Also, it is obvious that item #24 is relatively easier than the rest of the items.

There is another trend in the grade 5 ICCs. The female subgroup outperforms the male subgroup on most of the items for high ability levels. Additionally, in the most difficult items, (#18, 20, 22 and 23), the females outperform the males more dramatically than in the other items. This trend supports the findings of Marion and Shepard (1995) on this same data set. Performance assessments should favor females because their format is more context based and emphasizes writing more than multiple-choice tests do.

Conclusions

The first conclusion that can be reached from this data is that there is substantial DIF on certain items. The range of DIF indices has been discussed in the results section of this paper. In the grade 5 data, items 18, 23, 25 and 26 exhibit the largest DIF. Item 18 involves proofreading a letter. Item 23 is a written description comparing two samples in a salinity experiment. Item 25 involves writing a prediction for the

behavior of certain materials in a salinity experiment. Item 26 involves the recording of observations and measurements from another salinity experiment.

An examination of the actual items that exhibit large DIF in the 5th grade sample can illuminate some of the previously mentioned findings on gender differences in assessment. Females perform better on each of these items. None of these items involve algorithms. Consequently, our study does not support the conclusions of Berberoglu, 1995, etc..

It is impossible to determine interest level with the data available for this study. However, three of the four items that display large DIF are part of the salinity section of the assessment. Perhaps the topic of salinity is differentially interesting to males and females. Consequently, the work of Westers and Kelderman (1992) cannot be corroborated but may find support from these findings.

The context factor as a source of gender differences in assessment gains some credibility from this study. The four items on the grade 5 assessment that exhibit large DIF are all embedded in real contexts. In particular, item 18 involves a letter witting activity that also has some applicability outside of school. However, its DIF is similar to the other three items. Additionally, performance assessments themselves are more embedded in context than traditional multiple choice tests. Thus, we should and indeed do see females perform better across all items with DIF in this assessment.

The item format factor in gender differences in assessment gains considerable support from the results of this study. Each of the four items with large DIF involve considerable writing. This finding supports the work of Jovanovic et al., (1994) and Marion and Shepard, (1995).

The visual/spatial orientation factor does not receive corroboration from this study. The four items with large DIF do not involve visual/spatial orientation. Consequently, the previous research that suggests that gender DIF is due to the three factors of interest, embedded context and time format is supported by the results of this study. Other proposed factors in gender DIF do not receive corroboration from the findings of this study.

The second conclusion that can be reached from the results of this project is that there is considerably less DIF in the fifth grade science items than there is in the eighth grade science items. The DIF indices for the eighth grade range from .02 to .56. This is in comparison to the range of from .02 to .33 in the fifth grade items. This conclusion is logical when one considers the nature of gender differences in students in grade 5

as compared to students in grade 8. Adolescence affects eighth grade students more strongly than fifth grade students and may, in fact, emphasize gender differences because of the development of secondary sexual characteristics.

Based upon the current findings, we can answer our two research questions in the following manner. DIF due to gender differences exists in some of the items of the State Performance Assessment of 1994. Additionally, the data from this study support some of the factors that have been proposed by previous research to account for such DIF. The values of DIF for both of these sets of data range from .03 to .56. The DIF data from this project uniformly indicate that females do better on science performance assessments than do males. Also, the DIF in the fifth grade sample on this assessment is less than the DIF in the eighth grade sample.

There are several extensions of the current project that should be considered for future research. First, the problem with multidimensionality and the lack of local independence in this measure should be addressed by the application of the procedures recommended by Douglas, Roussos and Stout (1996). Second, the grade three sample of this assessment should be subjected to DIF analysis in order to provide a comparison for the current findings. Third, this assessment should be compared to other performance assessments and to more traditional forms of assessment. The two current trends in measurement toward both more "authentic" performance assessment and more stringent accountability measures makes the assumptions of item response theory and the differential item functioning of current assessments a critical issue. Comparison studies may serve to enlighten the measurement community in regard to these two trends.

TABLE 1:

Indices of DIF as well as the a and b parameter estimates by gender for 30 items on the 1994 5th grade — State Performance Assessment Program

item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
1	girls	0.91	-1.25			0.16		
	boys	0.98	-0.85					
	difference	0.07	-0.4					
2	girls	1.02	-0.56			0.22		
	boys	1.17	-0.05					
	difference	0.15	-0.51					
3	girls	1.24	-0.11	1.38		0.27	0.19	
	boys	1.42	0.45	1.78				
	difference	0.18	0.56	0.4				
4	girls	1.23	-0.14	1.88		0.26	0.16	
	boys	1.51	0.38	2.17				
	difference	0.28	0.52	0.29				
5	girls	1.19	-0.16	2.25		0.26	0.11	
	boys	1.42	0.37	2.43				
	difference	0.23	0.53	0.18				
6	girls	0.82	-0.37	2.36		0.2	0.1	
	boys	0.91	0.14	2.59				
	difference	0.09	0.51	0.23				
7	girls	1.18	0.13	4.03		0.13	0.03	
	boys	1.24	0.41	4.14				
	difference	0.06	0.28	0.11				
8	girls	0.71	0.21			0.2		
	boys	0.83	0.72					
	difference	0.12	0.51					
9	girls	0.7	-0.04			0.13		
	boys	0.8	0.29					
	difference	0.1	0.33					
10	girls	1.11	1.45			0.07		
	boys	1.31	1.4					
	difference	0.2	-0.05					
11	girls	0.94	1.21			0.12		
	boys	1.25	1.21					
	difference	0.31	0					

item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
12	girls	1.17	1.65			0.08		
	boys	1.27	1.48					
	difference	0.1	-0.17					
13	girls	0.69	1.44	2.46		0.03	0.03	
	boys	0.75	1.4	2.36				
	difference	0.06	-0.04	-0.1				
14	girls	0.74	-2.33	1.31	2.13	0.14	0.06	0.05
	boys	0.86	-1.85	1.3	2.07			
	difference	0.12	-0.48	-0.01	-0.06			
15	girls	0.87	0.7			0.12		
	boys	1.04	0.92					
	difference	0.17	0.22					
16	girls	1.06	0.38			0.14		
	boys	1.3	0.63					
	difference	0.24	0.25					
17	girls	0.97	-0.36	0.77		0.17	0.12	
	boys	1.27	-0.05	0.85				
	difference	0.3	-0.31	0.08				
18	girls	0.94	-3.2	0.74		0.32	0.11	
	boys	1.11	-2.19	0.95				
	difference	0.17	-1.01	0.21				
19	girls	0.56	-8.02	3		0.13	0.17	
	boys	0.91	-4.5	-4.5				
	difference	0.35	-3.52	-0.69				
20	girls	0.54	-7.89	3.28		0.15	0.15	
	boys	0.88	-4.38	-4.38				
	difference	0.34	-3.51	-0.76				
21	girls	0.87	1.85	4.22		0.09	0.02	
	boys	1.04	1.96	1.96				
	difference	0.17	0.11	-0.25				
22	girls	0.67	-2.1			0.3		
	boys	0.94	-1.19					
	difference	0.27	-0.91					
23	girls	0.5	-2	5.64		0.33	0.05	
	boys	0.75	-0.84	4.41				
	difference	0.25	-1.16	-1.23				

item #		a's	b-1's	b-2's	b-3's	b-1's DIF	b-2's DIF	b-3's DIF
24	girls	0.64	1.23			0.15		
	boys	0.9	1.05					
	difference	0.26	-0.18					
25	girls	1.01	-1.75	2.42		0.31	0.11	
	boys	1.32	-1.05	2.16				
	difference	0.31	-0.7	-0.26				
26	girls	0.58	-1.82	4.04		0.32	0.09	
	boys	0.85	-0.81	3.29				
	difference	0.27	-1.01	-0.75				
27	girls	0.91	0.14	3.74		0.21	0.06	
	boys	1.19	0.56	3.31				
	difference	0.28	0.42	-0.43				
28	girls	0.92	-1	4.84		0.22	0.08	
	boys	1.2	-0.52	3.92				
	difference	0.28	-0.48	-0.92				
29	girls	1.29	0.09	1.45		0.1	0.07	
	boys	1.57	0.21	1.41				
	difference	0.28	0.12	-0.04				
30	girls	1.18	0.74	3.4		0.12	0.02	
	boys	1.36	0.97	3.27				
	difference	0.18	0.23	-0.13				

Table 7

Various indices of DIF as well as the *a* and *b* parameter estimates by gender for 26 items (one cluster/form) on the science test from the 1994 eighth grade — State Performance Assessment Program.

Item	Gender	*a*	b_1	b_2	b_3	b_1DIF	b_2DIF	b_3DIF
7	Female	2.43	-1.32	-0.76				
7	Male	2.37	-1.35	-0.77				
	difference[2]	-0.06	-0.03	-0.01		**0.04**	**0.02**	
8	Female	2.12	-0.46					
8	Male	2.15	-0.50					
	difference	0.03	-0.04			**0.06**		
10	Female	1.71	-0.16	0.04	1.24			
10	Male	1.75	-0.12	0.12	1.33			
	difference	0.04	0.04	0.08	0.09	**0.06**	**0.22**	**0.13**
12	Female	1.85	0.81					
12	Male	1.98	0.83					
	difference	0.13	0.02			**0.07**		
18	Female	1.68	0.64	1.29				
18	Male	1.75	0.68	1.27				
	difference	0.07	0.04	-0.02		**0.07**	**0.05**	
4	Female	1.36	0.25					
4	Male	1.28	0.22					
	difference	-0.08	-0.03			**0.09**		
9	Female	1.77	0.43					
9	Male	1.63	0.42					
	difference	-0.14	-0.01			**0.09**		
15	Female	1.62	1.15					
15	Male	1.71	1.04					
	difference	0.09	-0.11			**0.09**		
6	Female	1.35	-0.01	1.77				
6	Male	1.44	0.03	1.58				
	difference	0.09	0.04	-0.19		**0.10**	**0.22**	
16	Female	1.90	1.28					
16	Male	2.08	1.22					
	difference	0.18	-0.06			**0.11**		

Table 2 (continued)

Item	Gender	*a*	b_1	b_2	b_3	b_1DIF	b_2DIF	b_3DIF
1	Female	1.22	-0.64					
1	Male	1.12	-0.62					
	difference	-0.10	0.02			**0.12**		
19	Female	1.97	1.61					
19	Male	1.92	1.52					
	difference	-0.05	-0.09			**0.13**		
17	Female	1.77	0.68	2.20				
17	Male	1.92	0.78	2.11				
	difference	0.15	0.10	-0.09		**0.15**	**0.11**	
23	Female	1.69	0.01	1.05				
23	Male	1.65	0.10	1.23				
	difference	-0.04	0.09	0.18		**0.15**	**0.24**	
24	Female	1.17	1.04					
24	Male	1.21	1.16					
	difference	0.04	0.12			**0.17**		
26	Female	1.31	0.37					
26	Male	1.19	0.48					
	difference	-0.12	0.11			**0.18**		
3	Female	1.49	0.36					
3	Male	1.29	0.38					
	difference	-0.20	0.02			**0.19**		
25	Female	1.23	0.98					
25	Male	1.25	1.14					
	difference	0.02	0.16			**0.22**		
21	Female	2.34	1.27	2.00				
21	Male	2.26	1.09	1.84				
	difference	-0.08	-0.18	-0.16		**0.25**	**0.22**	
11	Female	1.99	0.64					
11	Male	2.15	0.71					
	difference	0.16	0.07			**0.30**		
14	Female	1.41	1.08	1.75				
14	Male	1.51	0.85	1.49				
	difference	0.10	-0.23			**0.30**	**0.32**	

Table 2 (continued)

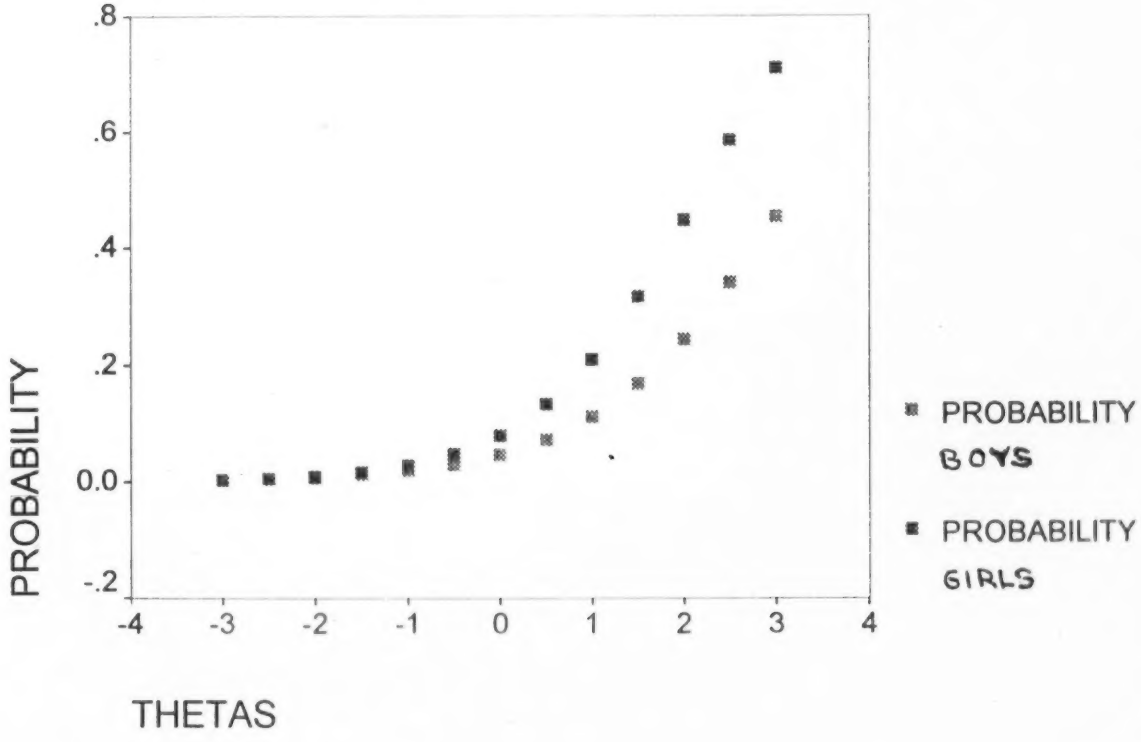
Item	Gender	a	b_1	b_2	b_3	b_1 DIF	b_2 DIF	b_3 DIF
27	Female	1.26	-1.04	0.85				
27	Male	1.34	-0.80	0.95				
	difference	0.08	0.24	0.10		**0.31**	**0.15**	
5	Female	1.22	-0.20	1.36				
5	Male	1.17	0.04	1.66				
	difference	-0.05	0.24	0.30		**0.33**	**0.36**	
28	Female	1.40	-0.51	2.23				
28	Male	1.40	-0.19	2.68				
	difference	0.00	0.32	0.45		**0.44**	**0.48**	
22	Female	1.53	0.46	1.82				
22	Male	1.66	0.13	1.40				
	difference	0.13	-0.33	-0.42		**0.46**	**0.54**	
20	Female	1.57	0.89					
20	Male	1.80	0.48					
	difference	0.23	-0.41			**0.56**		

Appendix

A

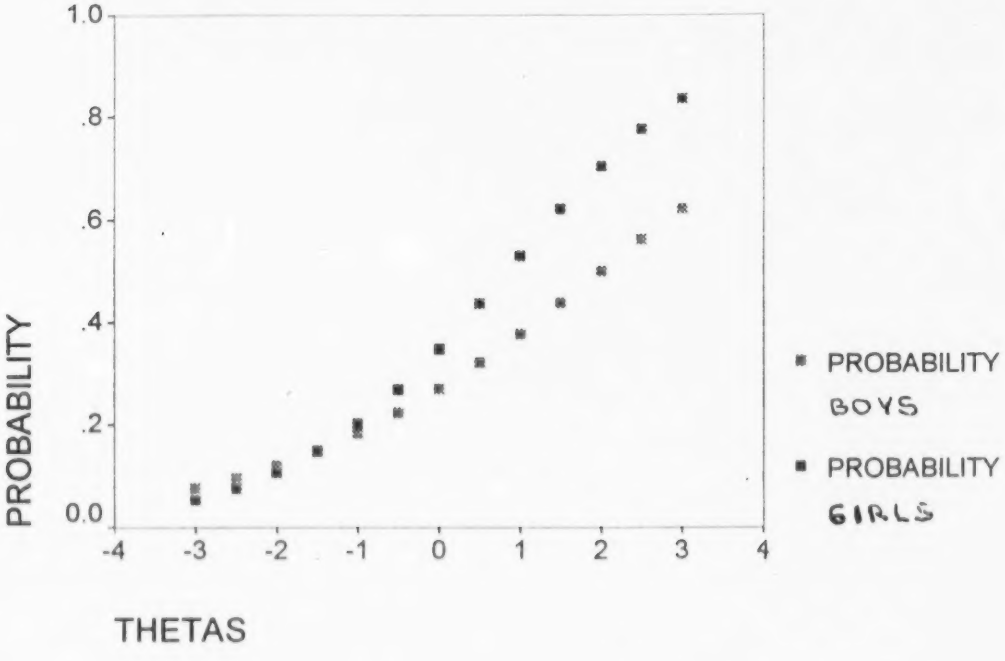
ICC ITEM 18

b1 dif .32



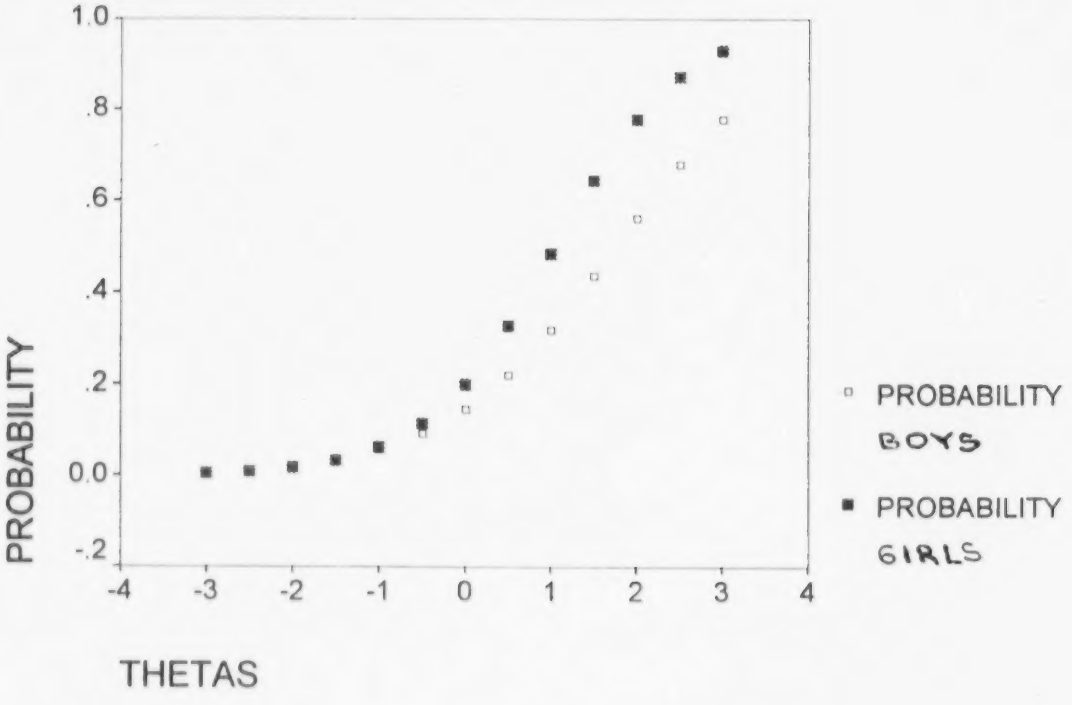
ICC ITEM 23

b1 dif .33



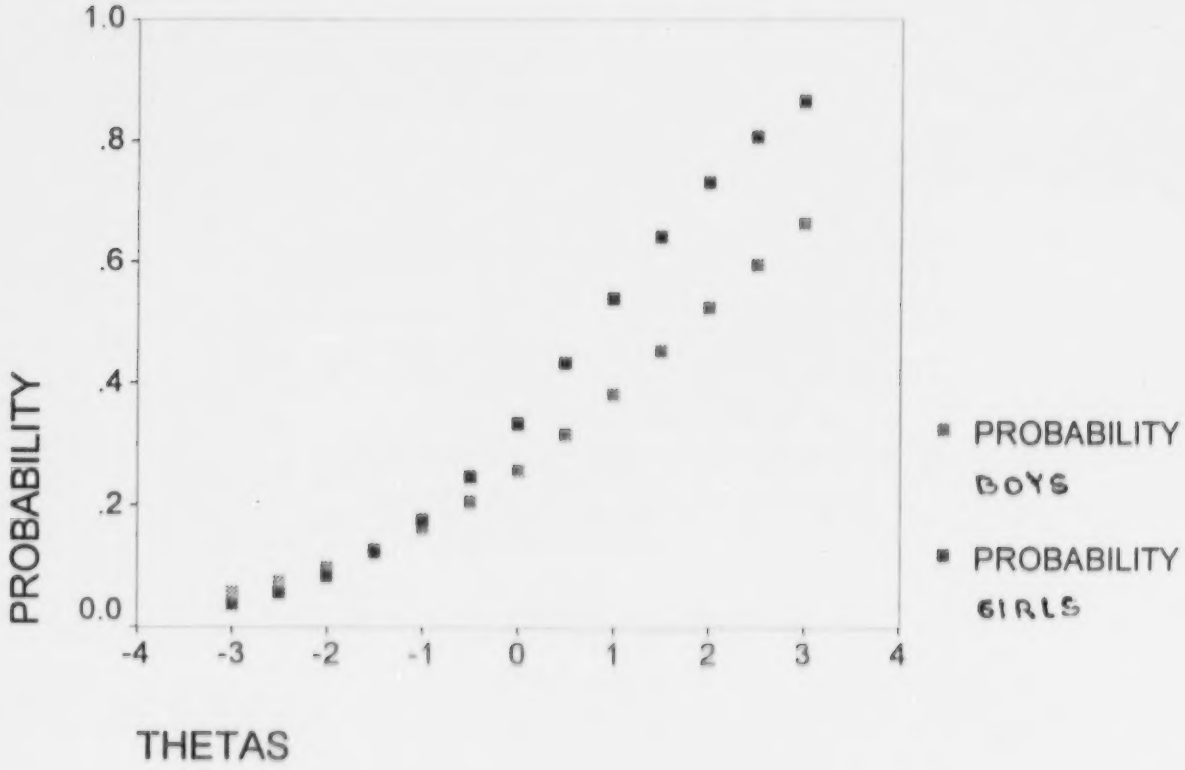
ICC ITEM 25

b1 dif .31



ICC ITEM 26

b1 dif .32



Reference List

1. Angoff, W. H. (1972, September). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association. Honolulu, HA. (ERIC Document Reproduction Service No. ED 069 686).

2. Berberoglu, G. (1995). Differential item functioning (DIF) analysis of computation, word problems and geometry questions across gender and SES groups. Studies in Educational Evaluation, 21, (4), 439-456.

3. Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Education Measurement, 27, 165-174.

4. Bransky, J. & Qualter, A. (1993). Applying physics concepts -- Uncovering the gender differences in assessment of performance unit results. Research in Science and Technological Education, 11, (2), 141-156.

5. Breland, H. M. (1991). A study of gender and performance on the Advanced Placement History Examination. College Board Report No. 91-4. New York: College Entrance Examination Board.

6. Bridgeman B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. Journal of Educational Measurement, 31, 1-6.

7. Camilli, G. & Shepard, L. A. (1994). Methods for identifying biased test items. Volume 4. Thousand Oaks, CA: Sage Publications.

8. Carlton, S. T. & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. Educational Testing Service. Princeton, NJ. (ERIC Document Reproduction Service No. ED 385 574).

9. Catsambis, S. (1994). The path to math: gender and racial-ethnic differences in mathematics participation from middle school to high school. Sociology of Education, 67, (3), 199-215.

10. Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: American Council on Education, MacMillan Publishing.

11. Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

12. Douglas, J. A., Roussos, L. A. & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. Journal of Educational Measurement, 33, (4), 465-484.

13. Friedman, L. (1996). Meta-analysis and quantitative gender differences: Reconciliation. Focus on Learning Problems in Mathematics, 18, (3), 123-128.

14. Gallagher, A. M. (1992). Sex differences in problem-solving strategies used by high-scoring examinees on the SAT-M. College Entrance Examination Board. New York, NY. (ERIC Document Reproduction Service No. ED 352 420).

15. Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

16. Jackson, L. A. et al. (1995). The numbers game: Gender and attention to numerical information. Sex Roles: A Journal of Research, 33, (7-8), 559-568.

17. Jovanovic, J. et al. (1994). Performance-based assessment: Will gender differences in science achievement be eliminated? Education and Urban Society, 26, (4), 352-366.

18. Lane, S. et al. (1995, April). Gender-related differential item functioning on a middle-school mathematics performance assessment. Paper presented at the Annual

Meeting of the American Educational Research Association. San Francisco, CA. (ERIC Document Reproduction Service No. ED 292 821).

19. Lawrence, I. M. & Curley, W. E. (1989). Differential item functioning for males and females on SAT - Verbal reading subscore items: Follow-up study. Educational Testing Service. Princeton, NJ. (ERIC Document Reproduction Service No. ED 395 969).

20. Linn, M. C. & Hyde, J. S. (1989). Gender, mathematics, and science. Educational Researcher, 18, (8), 17-19, 22-27.

21. Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis, 15, 1-16.

22. Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, (8), 15-20.

23. Marion, S. F. & Shepard, L. A. (1995, April). Equity issues for large-scale science performance assessments: an analysis of gender and race patterns. Paper presented at the annual meeting of the National Association for Research in Science Teaching. San Francisco, CA.

24. Ryan, K. E. & Chiu, S. (1996, April). Detecting DIF on mathematics items: the case for gender and calculator sensitivity. Paper presented at the Annual Meeting of the American Education Research Association. New York, NY. (ERIC Document Reproduction Service No. ED 395 998).

25. Severiens, S. E., Ten, D. & Geert, T. N. (1994). Gender differences in learning styles: A narrative review and quantitative meta-analysis. Higher Education, 27, (4), 487-501.

26. Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21, (4), 22-27.

27. Thissen, D. (1991). Multilog user's guide, Version 6.0. Chicago, IL: Scientific Software Inc..

28. Wang, N. & Lane, S. (1994, April). Detection of gender-based differential item functioning in a mathematics performance assessment. Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA. (ERIC Document Reproduction Service No. ED 377 247).

29. Westers, P. & Kelderman, H. (1992). Examining differential item functioning due to item difficulty and alternative attractiveness. Psychometrika, 57, (1), 107-118.

30. Yen, W. M., et al. (1992). Final technical report: Maryland School Performance Assessment Program 1991. Monterey CA: CTB Macmillan/McGraw-Hill.

31. ----- (1996?). KIRIS Technical Report. Boulder, CO: Bob Linn.

32. ----- (1990). Learning outcomes in mathematics, reading, writing/language usage, social studies, and science. Maryland School Performance Assessment Program. Baltimore MD: Author.

33. ----- (1991). 386 - Multilog 6. Scientific Software, Inc.. Chicago, IL: Author.